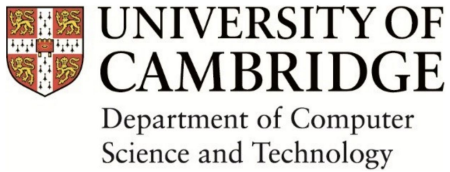
Federated Continual Learning with Adaptive Fuzzy Tiling Activation

Chenyang Ma (cm2196)
Yuwei Zhang (yz798)



Federated Learning (L361)
Project Presentation

Federated Continual Learning (FCL)

- Data on devices evolve over time: users delete, transfer, and upload data
- Model trains on a sequence of T tasks $\{\mathcal{T}^{(1)}, \mathcal{T}^{(2)}, ..., \mathcal{T}^{(T)}\}$
  - without access to training data of the previous tasks

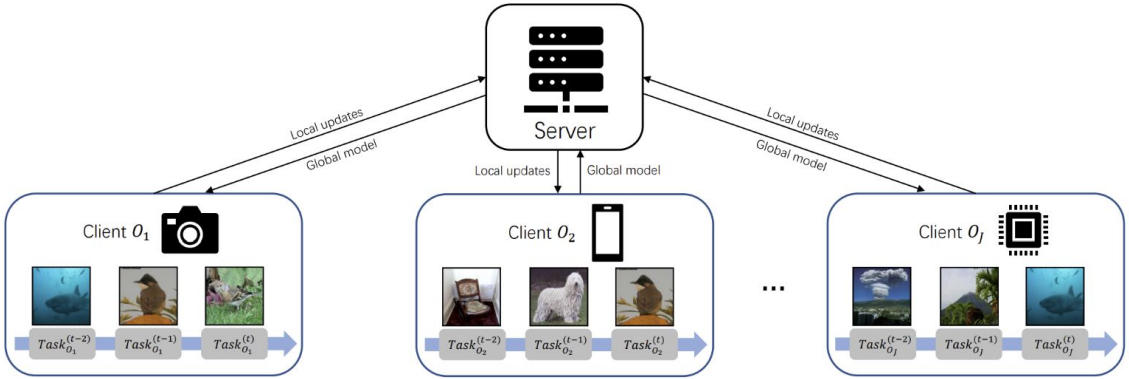
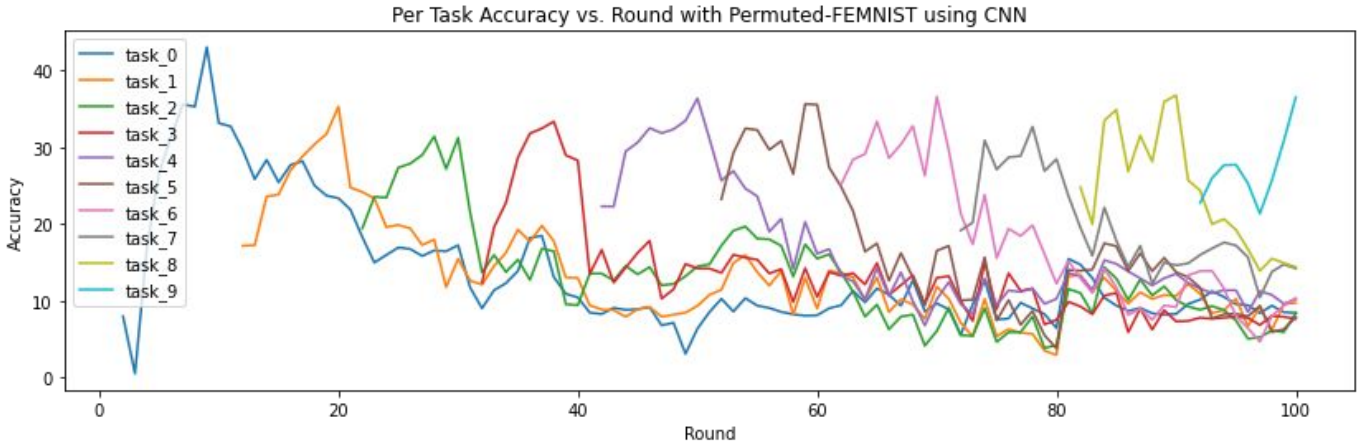


Figure 1: Federated continual learning. Each client is trained on task sequences. The server aggregates local model updates and broadcast the global model to all clients.

Catastrophic Forgetting in CL & FCL

- FCL inherits ***Catastrophic Forgetting*** from continual learning (CL) [1, 2]
- Knowledge from the new task overwrites the knowledge acquired from previous tasks: ***Interference***



[1] Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang. Federated continual learning with weighted inter-client transfer. In ICML, 2021.
[2] Yatin Chaudhary, Pranav Rai, Matthias Schubert, Hinrich Schütze, and Pankaj Gupta. Federated continual learning for text classification via selective inter-client transfer. In EMNLP, 2022.

Goal of this work

- Previous works on FCL develop FL protocols that are task-specific
  - Yoon et al. [1]: Image classification
  - Chaudhary et al. [2]: NLP

- Can we acquire a generalized solution to tackle ***Catastrophic Forgetting*** for FCL by applying techniques that are effective in CL?

[1] Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang. Federated continual learning with weighted inter-client transfer. In ICML, 2021.
[2] Yatin Chaudhary, Pranav Rai, Matthias Schubert, Hinrich Schütze, and Pankaj Gupta. Federated continual learning for text classification via selective inter-client transfer. In EMNLP, 2022.

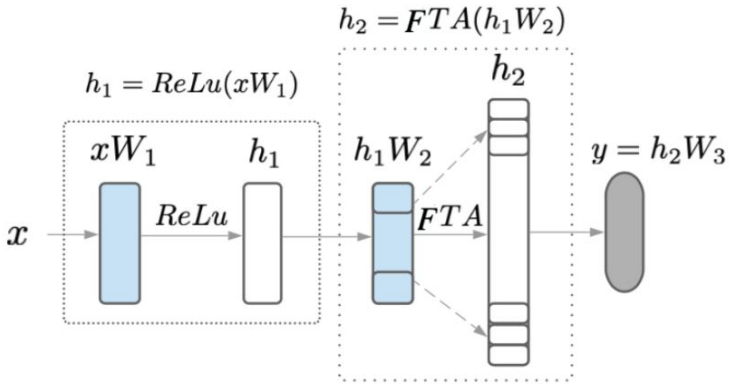
Natural Sparsity: Fuzzy Tiling Activation (FTA)

- Introduces sparsity by design: mapping a scalar to a vector [1]
- Task agnostic
- Model agnostic

$$v := (l, l+\xi, l+2\xi, ..., u-2\xi, u-\xi, u) = \text{arange}(l, u+\epsilon, \xi)$$

$$I_+(w) := \begin{cases} 1, & \text{if } w > 0 \\ 0, & \text{otherwise} \end{cases}$$

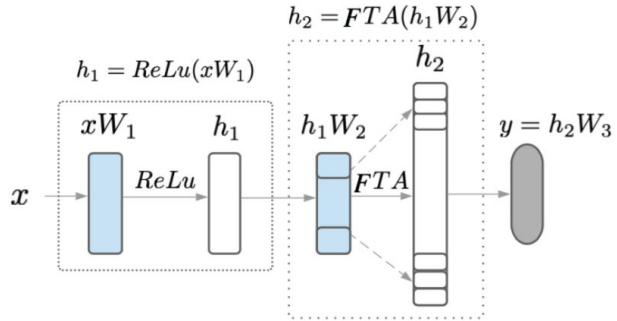
$$\vartheta(z) := 1 - I_+ \left(\text{ReLU}(v - z) + \text{ReLU}(z - \xi - v)\right)$$

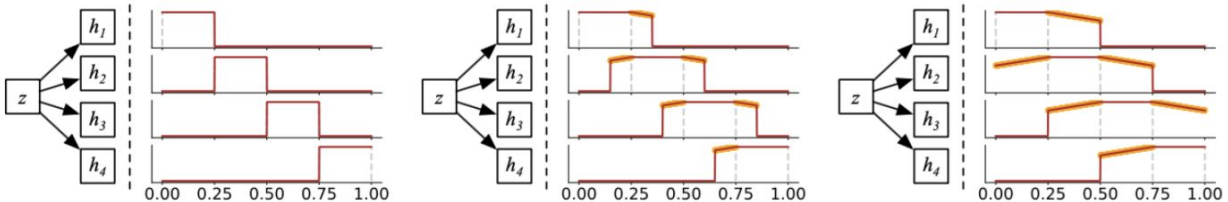


[1] Yangchen Pan, Kirby Banman, and Martha White. Fuzzy tiling activations: A simple approach to learning sparse representations online. In ICLR, 2021.

Natural Sparsity: Fuzzy Tiling Activation (FTA)

- Maps an input scalar to a smoothed one-hot encoding
- Leaks gradients
- Creates more sparse feature space





(a) TA, $k = 4$ (b) FTA, $k = 4, \eta = 0.1$ (c) FTA, $k = 4, \eta = 0.25$

$$I_{\eta,+}(w) := I_+(\eta - w) \times w + I_+(w - \eta) = \begin{cases} w, & \text{if } w < \eta \\ 1, & \text{otherwise} \end{cases}$$

$$\vartheta_\eta(z) := 1 - I_{\eta,+}\left(\mathrm{ReLU}(v - z) + \mathrm{ReLU}(z - \xi - v)\right)$$

[1] Yangchen Pan, Kirby Banman, and Martha White. Fuzzy tiling activations: A simple approach to learning sparse representations online. In ICLR, 2021.

Proposed method: Adaptive Fuzzy Tiling Activation (AFTA)

- FTA leads to ***gradient vanishing*** if the input scalar is not within the predefined bounds
- We argue the problem worsens in FL with data heterogeneity
- We develop a method to adaptively set the lower and upper bounds during training, named as **AFTA**

$$\forall q \in \{1, 2, ..., Q\}, h_q \in S$$

$$l = \min\{l \in S : l > \mu + 2\sigma\}, \quad u = \max\{u \in S : u < \mu - 2\sigma\}$$

In Combination with Continual Learning Techniques

- Knowledge Distillation (KD) [1]
  - Teacher model: global model trained on t-1 previous tasks
  - Student model: train on the t-th task locally

$$\mathcal{L}_{\mathrm{KD}} = (1-\alpha)\mathcal{L}_{\mathrm{CE}} + \alpha\mathcal{L}_{\mathrm{KL}}$$

$$\mathcal{L}_{\mathrm{KD}} = (1-\alpha)\left(-\sum_i y_i \log(p_{s_i})\right) + \alpha\left(\sum_i p_{t_i}^T \log\left(\frac{p_{t_i}^T}{p_{s_i}^T}\right)\right)$$

- Elastic Weight Consolidation (EWC) [2]
  - Regularize the network's parameters based on their importance for the previous tasks

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i}),$$

[1] Zhizhong Li and Derek Hoiem. Learning without forgetting. In ECCV, 2016.
[2] Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. arXiv, 2016.

Pipeline

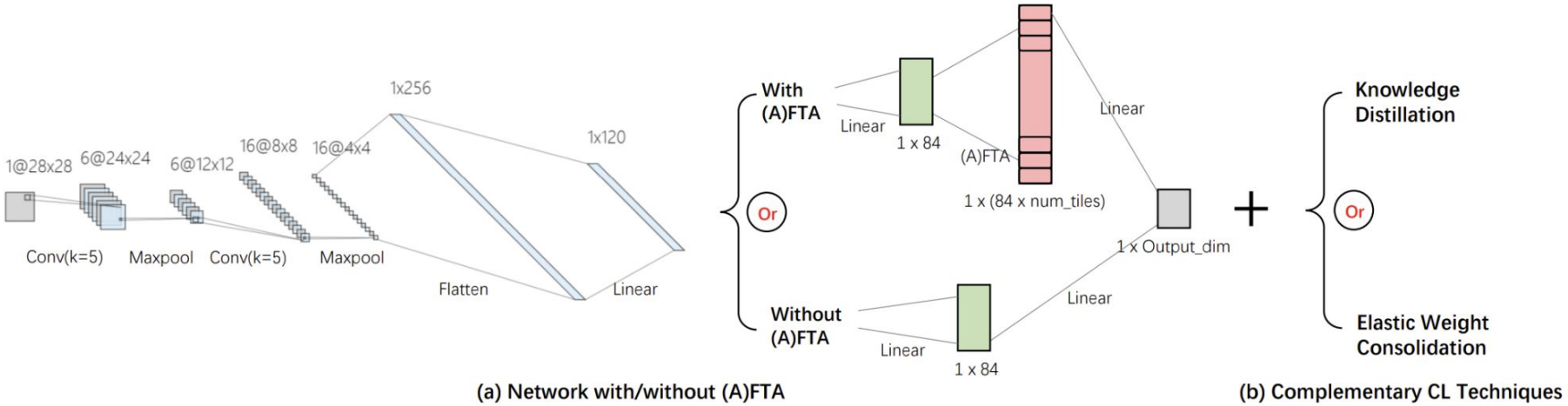
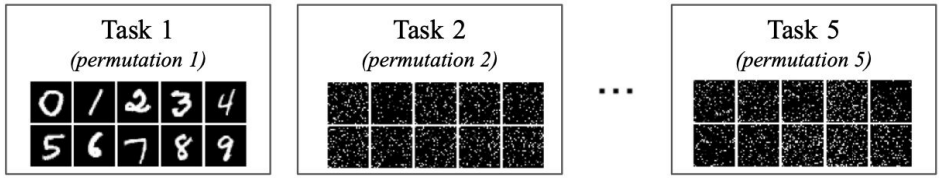


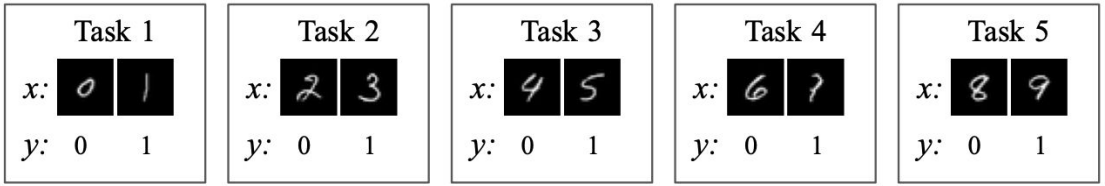
Figure 4: The pipeline. (a) shows the network with/without (A)FTA on the client-side; (b) shows the complementary CL techniques that can be combined with (A)FTA, which is also deployed on the client-side.

Experimental Design

- Permuted-FEMNIST
  - Each task is a ten-way classification
  - A different random pixel-level permutation is applied to the input data X



- Split-FEMNIST
  - Each task is a two-way classification
  - The original FEMNIST dataset for each client is split



Experimental Design

- Evaluation metrics
  - The final accuracy of all trained tasks
  - The best accuracy achieved by each task
  - Averaged incremental accuracy [1, 2, 3]
  - Averaged forgetting [1, 2, 3]

[1] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In CVPR, 2017.
[2] Haiyan Yin, Ping Li, et al. Mitigating forgetting in online continual learning with neuron calibration. In NeurIPS, 2021.
[3] Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang. Federated continual learning with weighted inter-client transfer. In ICML, 2021.

Results

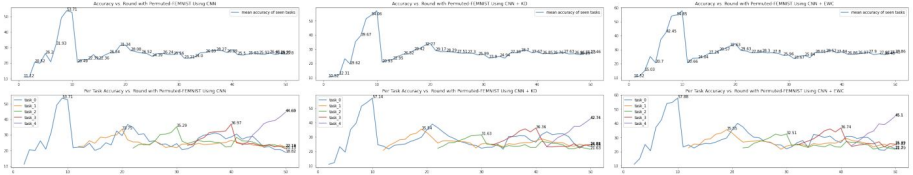
Table 1: Performances on Permuted-FEMNIST. **Accuracy** (↑), **Forgetting** ↓. The baseline is in **bold**. The highest Accuracy is in **red**. The lowest forgetting is in **blue**.

Sparsity	CL Tech	$\mathcal{T}^{(1)}$	$\mathcal{T}^{(2)}$	$\mathcal{T}^{(3)}$	$\mathcal{T}^{(4)}$	$\mathcal{T}^{(5)}$	$\mathcal{T}^{(1)*}$	$\mathcal{T}^{(2)*}$	$\mathcal{T}^{(3)*}$	$\mathcal{T}^{(4)*}$	$\mathcal{T}^{(5)*}$	Accuracy	Forgetting
None	None	**18.82**	**22.18**	**22.19**	**21.13**	**44.69**	**53.71**	**33.75**	**35.29**	**36.97**	**44.69**	**25.91**	**15.08**
	KD	23.55	24.58	21.63	24.81	42.74	57.14	35.84	31.63	36.36	42.74	26.96	**13.28**
	EWC	22.20	24.82	21.79	25.39	45.10	57.88	35.55	32.51	36.74	45.10	27.26	13.70
FTA	None	18.79	17.7	18.56	31.51	56.64	76.79	56.84	49.41	56.64	56.64	28.7	30.62
	KD	16.77	19.99	20.36	29.55	54.72	69.37	50.08	41.33	54.42	54.72	27.56	25.71
	EWC	28.29	22.94	23.06	36.56	59.37	79.59	58.49	50.05	66.91	59.37	34.29	28.84
AFTA	None	20.05	25.84	20.77	31.08	46.17	79.06	54.03	50.6	61.82	46.17	30.21	29.55
	KD	28.70	28.45	29.83	36.57	54.35	80.63	47.19	45.74	49.13	54.35	**34.48**	19.83
	EWC	28.56	28.74	29.62	35.65	54.59	80.90	47.68	44.88	49.84	54.59	34.42	20.15

Table 2: Performances on Split-FEMNIST. **Accuracy** (↑), **Forgetting** ↓. The baseline is in **bold**. The highest Accuracy is in **red**. The lowest forgetting is in **blue**.

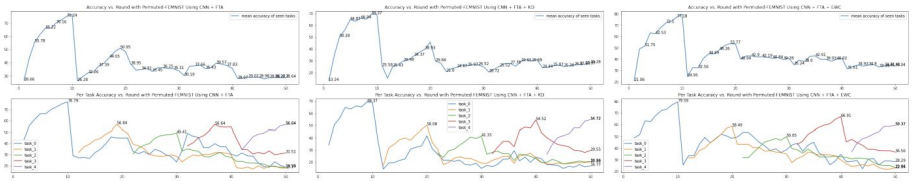
Sparsity	CL Tech	$\mathcal{T}^{(1)}$	$\mathcal{T}^{(2)}$	$\mathcal{T}^{(3)}$	$\mathcal{T}^{(4)}$	$\mathcal{T}^{(5)}$	$\mathcal{T}^{(1)*}$	$\mathcal{T}^{(2)*}$	$\mathcal{T}^{(3)*}$	$\mathcal{T}^{(4)*}$	$\mathcal{T}^{(5)*}$	Accuracy	Forgetting
None	None	**59.11**	**74.58**	**54.27**	**97.05**	**84.06**	**95.08**	**86.63**	**90.84**	**99.23**	**84.06**	**72.94**	**17.35**
	KD	74.02	61.45	46.95	96.62	85.17	97.56	88.88	70.01	99.52	85.17	75.69	15.39
	EWC	81.45	85.91	51.25	97.68	86.14	96.93	92.25	89.38	98.89	86.14	78.58	12.23
FTA	None	86.66	61.57	44.68	96.42	86.61	98.12	82.90	85.77	99.03	86.61	75.19	15.30
	KD	94.63	69.47	41.78	97.15	87.22	97.51	83.68	60.77	98.28	87.22	78.25	7.44
	EWC	84.15	72.83	45.37	95.55	88.17	97.42	89.01	78.30	98.82	88.17	78.70	13.13
AFTA	None	89.85	81.55	53.2	96.62	80.02	97.75	88.81	90.53	98.48	80.02	**82.68**	10.87
	KD	97.23	84.10	50.26	97.70	81.88	97.91	86.75	77.87	97.99	81.88	82.23	**6.25**
	EWC	86.35	70.40	47.44	96.37	86.48	97.80	87.14	78.74	99.01	86.48	78.83	12.43

Results: Full Training Curves



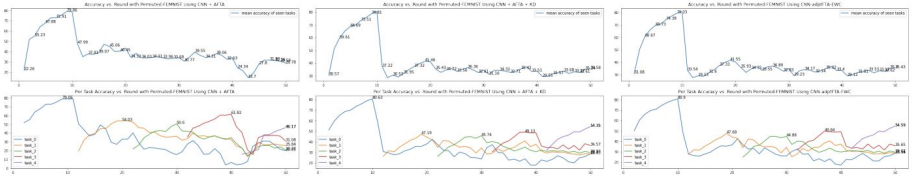
(a) CNN (b) CNN + KD (c) CNN + EWC

Figure 7: Performance on Permuted-FEMNIST Using vanilla CNN, (a) is baseline.



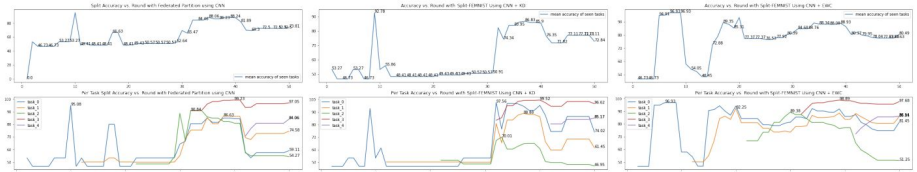
(a) CNN + FTA (b) CNN + FTA + KD (c) CNN + FTA + EWC

Figure 8: Performance on Permuted-FEMNIST Using CNN + FTA.



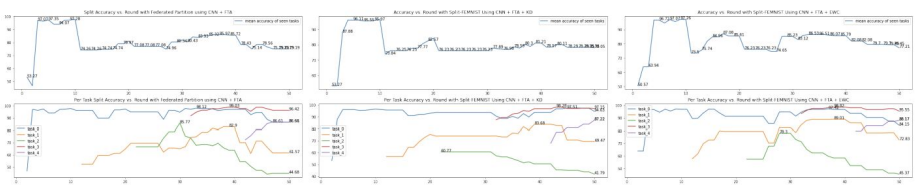
(a) CNN + AFTA (b) CNN + AFTA + KD (c) CNN + AFTA + EWC

Figure 9: Performance on Permuted-FEMNIST Using CNN + AFTA.



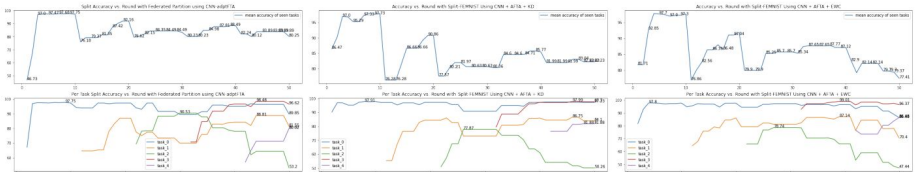
(a) CNN (b) CNN + KD (c) CNN + EWC

Figure 10: Performance on Split-FEMNIST Using vanilla CNN, (a) is baseline.



(a) CNN + FTA (b) CNN + FTA + KD (c) CNN + FTA + EWC

Figure 11: Performance on Split-FEMNIST Using CNN + FTA.



(a) CNN + AFTA (b) CNN + AFTA + KD (c) CNN + AFTA + EWC

Figure 12: Performance on Split-FEMNIST Using CNN + AFTA.

THANK YOU

UNIVERSITY OF CAMBRIDGE
Department of Computer Science and Technology

Federated Learning (L361)
Project Presentation